doi:10.1038/nature09042

LETTERS

Putting brain training to the test

Adrian M. Owen[1], Adam Hampshire[1], Jessica A. Grahn[1], Robert Stenton[2], Said Dajani[2], Alistair S. Burns[3], Robert J. Howard[2] & Clive G. Ballard[2]

'Brain training', or the goal of improved cognitive function through the regular use of computerized tests, is a multimillion-pound industry[1], yet in our view scientific evidence to support its efficacy is lacking. Modest effects have been reported in some studies of older individuals[2,3] and preschool children[4], and video-game players outperform non-players on some tests of visual attention[5]. However, the widely held belief that commercially available computerized brain-training programs improve general cognitive function in the wider population in our opinion lacks empirical support. The central question is not whether performance on cognitive tests can be improved by training, but rather, whether those benefits transfer to other untrained tasks or lead to any general improvement in the level of cognitive functioning. Here we report the results of a six-week online study in which 11,430 participants trained several times each week on cognitive tasks designed to improve reasoning, memory, planning, visuospatial skills and attention. Although improvements were observed in every one of the cognitive tasks that were trained, no evidence was found for transfer effects to untrained tasks, even when those tasks were cognitively closely related.

To investigate whether regular brain training leads to any improvement in cognitive function, viewers of the BBC popular science programme 'Bang Goes The Theory' participated in a six-week online study of brain training. An initial 'benchmarking' assessment included a broad neuropsychological battery of four tests that are sensitive to changes in cognitive function in health and disease[6–12]. Specifically, baseline measures of reasoning[6], verbal short-term memory (VSTM)[7,12], spatial working memory (SWM)[8–10] and paired-associates learning (PAL)[11,13] were acquired. Participants were then randomly assigned to one of two experimental groups or a third control group and logged on to the BBC Lab UK website to practise six training tasks for a minimum of 10 min a day, three times a week. In experimental group 1, the six training tasks emphasized reasoning, planning and problem-solving abilities. In experimental group 2, a broader range of cognitive functions was trained using tests of short-term memory, attention, visuospatial processing and mathematics similar to those commonly found in commercially available brain-training devices. The difficulty of the training tasks increased as the participants improved to continuously challenge their cognitive performance and maximize any benefits of training. The control group did not formally practise any specific cognitive tasks during their 'training' sessions, but answered obscure questions from six different categories using any available online resource. At six weeks, the benchmarking assessment was repeated and the pre- and post-training scores were compared. The difference in benchmarking scores provided the measure of generalized cognitive improvement resulting from training. Similarly, for each training task, the first and last scores were compared to give a measure of specific improvement on that task.

Of 52,617 participants aged 18–60 who initially registered, 11,430 completed both benchmarking assessments and at least two full training sessions during the six-week period. On average, participants completed 24.47 (s.d. = 16.95) training sessions (range = 1–188 sessions). The three groups were well matched in age (39.14 (11.91), 39.65 (11.83), 40.51 (11.79), respectively) and gender (female/male = 5.5:1, 5.6:1 and 4.3:1, respectively).

Numerically, experimental group 1 improved on four benchmarking tests and experimental group 2 improved on three benchmarking tests (Fig. 1), with standardized effect sizes varying from small (for example, 0.35 (99% confidence interval (CI), 0.29–0.41)) to very small (for example, 0.01 (99% CI, −0.05–0.07)). However, the control group also improved numerically on all four tests with similar effect sizes (Table 1). When the three groups were compared directly, effect sizes across all four benchmarking tests were very small (for example, 0.01 (99% CI, −0.05–0.07) to 0.22 (99% CI, 0.15–0.28)) (Table 2). In fact, for VSTM and PAL, the difference between benchmarking sessions was numerically greatest for the control group (Fig. 1, Table 1 and Table 2). These results suggest an equivalent and marginal test–retest practice effect in all groups across all four tasks (Table 1). In contrast, the improvement on the tests that were actually trained was convincing across all tasks for both experimental groups. For example, for the tasks practised by experimental group 1, differences were observed with large effect sizes of between 0.73 (99% CI, 0.68–0.79) and 1.63 (99% CI, 1.57–1.7) (Table 3 and Fig. 2). Using Cohen's[14] notion that 0.2 represents a small effect, 0.5 a medium effect and 0.8 a large effect, even the smallest of these improvements would be considered large. Similarly, for experimental group 2, large improvements were observed on all training tasks, with effect sizes of between 0.72 (99% CI, 0.67–0.78) and 0.97 (99% CI, 0.91–1.03) (Table 3 and Fig. 2). Numerically, the control group also improved in their ability to answer obscure knowledge questions, although the effect size was small (0.33 (99% CI, 0.26–0.4)) (Table 3 and Fig. 2). In all three groups, whether these improvements reflected the simple effects of task repetition (that is, practise), the adoption of new task strategies, or a combination of the two is unclear, but whatever the process effecting change, it did not generalize to the untrained benchmarking tests.

The relationship between the number of training sessions and changes in benchmark performance was negligible in all groups for all tests (largest Spearman's $\rho = 0.059$; Supplementary Fig. 1). The effect of age was also negligible (largest Spearman's $\rho = -0.073$). Only two tests showed a significant effect of gender (PAL in experimental group 1 and VSTM in experimental group 2), but the effect sizes were very small (0.09 (99% CI, −0.01–0.2) and 0.09 (99% CI, −0.03–0.2), respectively).

These results provide no evidence for any generalized improvements in cognitive function following brain training in a large sample of healthy adults. This was true for both the 'general cognitive training' group (experimental group 2) who practised tests of memory, attention, visuospatial processing and mathematics similar to many of those found in commercial brain trainers, and for a more focused

[1]MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 7EF, UK. [2]King's College London, Institute of Psychiatry, De Crespigny Park, London SE5 8AF, UK. [3]University of Manchester and Manchester Academic Health Science Centre, Manchester M13 9PL, UK.

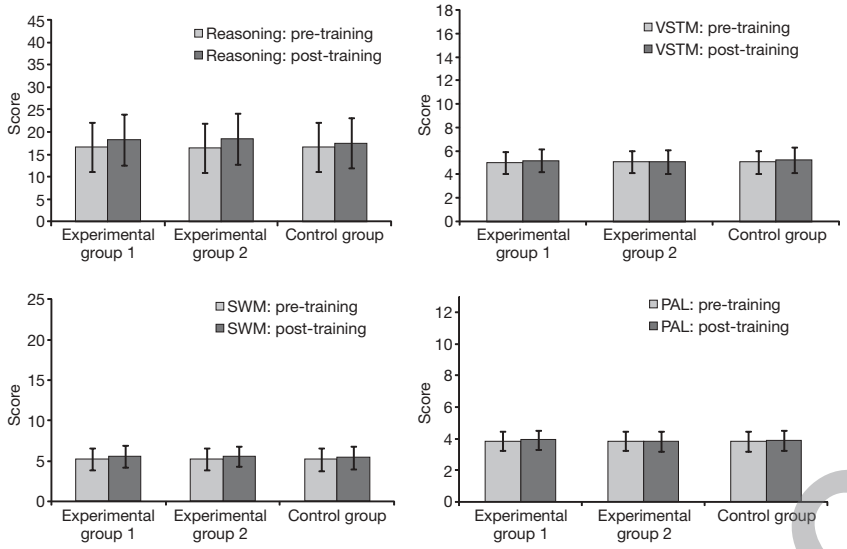

Figure 1 | Benchmarking scores at baseline and after six weeks of training across the three groups of participants. PAL, paired-associates learning; SWM, spatial working memory; VSTM, verbal short-term memory. Bars represent standard deviations.

training group (experimental group 1) who practised tests of reasoning, planning and problem solving. Indeed, both groups provided evidence that training-related improvements may not even generalize to other tasks that use similar cognitive functions. For example, three of the tests practised by experimental group 1 (reasoning 1, 2 and 3) specifically emphasized abstract reasoning abilities, yet numerically larger changes on the benchmarking test that also required abstract reasoning were observed in experimental group 2, who were not trained on any test that specifically emphasized reasoning. Similarly, of all the trained tasks, memory 2 (based on the classic parlour game in which players have to remember the locations of objects on cards) is most closely related to the PAL benchmarking task (in which participants also have to remember the locations of objects), yet numerically, PAL performance actually deteriorated in the experimental group that trained on the memory 2 task (Fig. 1).

Could it be that no generalized effects of brain training were observed because the wrong types of cognitive task were used? This is unlikely because 12 different tests, covering a broad range of cognitive functions, were trained in this study. In addition, the six training tasks that emphasized abstract reasoning, planning and problem solving were included specifically because such tasks are known to correlate highly with measures of general fluid intelligence or '*g*'[15–17], and were therefore most likely to produce an improvement in the general level of cognitive functioning. Indeed, functional neuroimaging studies have revealed clear overlap in frontal and parietal regions between similar tests of reasoning and planning to those used here[15,17–19] and tests that are specifically designed to measure *g*[15,20], whereas damage to the frontal lobe impairs performance on both types of task[10,16,21].

Is it possible that the benchmarking tests were insensitive to the generalized effects of brain training? This is also unlikely because the benchmarking tests were chosen for their known sensitivity to small changes in cognitive function in disease or following low-dose neuropharmacological interventions in healthy volunteers. For example,

Table 1 | Changes between pre- and post-training benchmarking performance for each group

Test	Measure	Experimental group 1	Experimental group 2	Control group
Reasoning	Mean difference	1.73	1.97	0.90
	Effect size	0.31	0.35	0.16
	99% CI	0.26–0.36	0.29–0.41	0.09–0.23
VSTM	Mean difference	0.15	0.03	0.22
	Effect size	0.16	0.03	0.21
	99% CI	0.11–0.21	−0.02–0.09	0.14–0.28
SWM	Mean difference	0.33	0.35	0.27
	Effect size	0.24	0.27	0.19
	99% CI	0.19–0.29	0.21–0.33	0.12–0.26
PAL	Mean difference	0.06	−0.01	0.07
	Effect size	0.10	0.01	0.11
	99% CI	0.05–0.16	−0.05–0.07	0.04–0.18

CI, confidence interval; PAL, paired-associates learning; SWM, spatial working memory; VSTM, verbal short-term memory.

Table 2 | Comparisons of each group's change in pre- and post-training benchmarking performance

Test	Measure	Experimental group 1 versus experimental group 2	Experimental group 1 versus control group	Experimental group 2 versus control group
Reasoning	Mean difference	−0.231	0.831	1.062
	Effect size	0.05	0.17	0.22
	99% CI	−0.01–0.1	0.1–0.23	0.15–0.28
VSTM	Mean difference	0.130	−0.056	−0.186
	Effect size	0.13	0.05	0.18
	99% CI	0.07–0.18	−0.01–0.12	0.11–0.24
SWM	Mean difference	−0.028	0.057	0.085
	Effect size	0.02	0.04	0.06
	99% CI	−0.04–0.07	−0.03–0.1	−0.01–0.12
PAL	Mean difference	0.117	−0.012	−0.129
	Effect size	0.10	0.01	0.11
	99% CI	0.04–0.15	−0.05–0.07	0.04–0.17

See Table 1 for definitions.

Table 3 | Changes between first and last training scores for each group

Experimental group	Test	Mean difference	Effect size	99% CI
Experimental group 1	Reasoning 1	33.96	1.63	1.57–1.7
	Reasoning 2	13.45	1.03	0.98–1.09
	Reasoning 3	11.45	1.25	1.19–1.31
	Planning 1	15.17	1.28	1.23–1.34
	Planning 2	14.42	1.10	1.05–1.16
	Planning 3	10.41	0.73	0.68–0.79
Experimental group 2	Maths	18.15	0.90	0.84–0.96
	Visuospatial	8.62	0.95	0.89–1.02
	Attention 1	9.71	0.93	0.87–0.99
	Attention 2	8.48	0.84	0.78–0.9
	Memory 1	7.29	0.72	0.67–0.78
	Memory 2	5.30	0.97	0.91–1.03
Control group	Questions	3.62	0.33	0.26–0.40

For description of tests, see Methods.

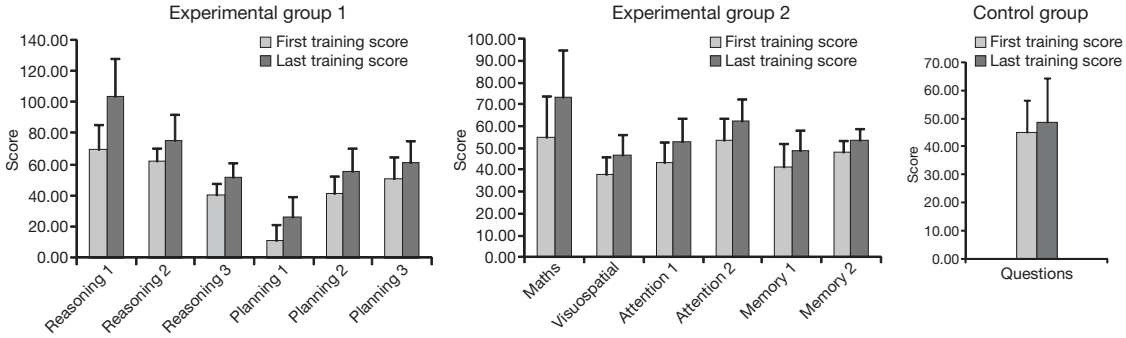

Figure 2 | First and last training scores for the six tests used to train experimental group 1 and experimental group 2. The first and last scores for the control group are also shown. Bars represent standard deviations.

the SWM task is sensitive to damage to the frontal cortex[10,22] and impairments are observed in patients with Parkinson's disease[23]. On the other hand, low-dose methylphenidate improves performance on the same task in healthy volunteers[8,9]. Similarly, the PAL task is highly sensitive to various neuropathological conditions, including Alzheimer's disease[11], Parkinson's disease[13] and schizophrenia[24], whereas the α2-agonists guanfacine and clonidine improve performance in healthy volunteers[25].

Could it be that improvements in the experimental groups were 'masked' by the direct comparison with the control group, who were, arguably, also exercising attention, planning and visuospatial processes? This seems unlikely because there was a clear difference between the substantial improvements in both experimental groups across all trained tasks and the very modest improvement observed in the control group on their obscure knowledge test, suggesting that the experimental groups did benefit more from their training programmes, albeit only on the tasks that were actually being trained. In any case, in all three groups the standardized effect sizes of the transfer effects were, at best, small (Table 1), suggesting that any comparison (even with a control group who did nothing) would have yielded a negligible brain training effect in the experimental groups.

It is important to consider the possibility that the amount of practise was insufficient to produce a measurable transfer effect of brain training in this study. Given the known sensitivity of the benchmarking tests[8–11,13,22–26], it seems reasonable to expect that 25 training sessions would yield a measurable group effect if one was present. More directly however, there was a negligible correlation between the number of training sessions and improvement in benchmarking scores (despite a strong correlation with improvement on training tasks; Supplementary Fig. 2), confirming that the amount of practise was unrelated to any generalized brain-training effect. That said, the possibility that an even more extensive training regime may have eventually produced an effect cannot be excluded.

To illustrate the size of the transfer effects observed in this study, consider the following representative example from the data. The increase in the number of digits that could be remembered following training on tests designed, at least in part, to improve memory (for example, in experimental group 2) was three-hundredths of a digit. Assuming a linear relationship between time spent training and improvement, it would take almost four years of training to remember one extra digit. Moreover, the control group improved by two-tenths of a digit, with no formal memory training at all.

In our view these results provide no evidence to support the widely held belief that the regular use of computerized brain trainers improves general cognitive functioning in healthy participants beyond those tasks that are actually being trained. Although we cannot exclude the possibility that more focused approaches, such as face-to-face cognitive training[2], may be beneficial in some circumstances, we believe that these results confirm that six weeks of regular computerized brain training confers no greater benefit than simply answering general knowledge questions using the internet.

METHODS SUMMARY

Of 11,430 participants who met the inclusion criteria, 4,678 were randomly assigned to experimental group 1, 4,014 to experimental group 2 and 2,738 to the control group. Over six weeks, experimental group 1 completed an average of 28.39 (s.d. = 19.86) training sessions, compared with 23.86 (15.66) in experimental group 2 and 18.66 (12.87) in the control group. All three groups were given the same four benchmarking tests (grammatical reasoning[6], VSTM[7,12], SWM[8–10], PAL[11,13]), immediately after registering for the trial and again six weeks later, irrespective of how many training or control sessions they had chosen to complete in between. The benchmarking tests were adapted from publicly available cognitive assessment tools designed and validated at the Medical Research Council Cognition and Brain Sciences Unit (by A.H. and A.M.O) and made freely available at http://www.cambridgebrainsciences.com. During the six-week training period the first experimental group was trained on six reasoning, planning and problem-solving tasks, while the second experimental group was trained on six tests of memory, attention, visuospatial processing and mathematical calculations, similar to those commonly found in commercially available brain-training programs. In each 'training' session, the control group was asked five obscure knowledge questions from one of six general categories and were asked to place answers in correct chronological order using any available online resource. The main outcome measures were the difference scores (post-training minus pre-training) for the four benchmarking tests in the two experimental groups and the control group. Changes in performance on the tests that were actually trained were also calculated by comparing the scores from the first and last training sessions. Owing to the large number of participants in this study, the size of any observed differences between the groups was quantified by reporting effect sizes[14] together with estimates of the likely margin of error (99% confidence intervals).

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 25 January; accepted 29 March 2010.
Published online 20 April 2010.

1. Aamodt, S. & Wang, A. Exercise on the brain. *The New York Times* ⟨http://www.nytimes.com/2007/11/08/opinion/08aamodt.html?_r=1⟩ (2007).
2. Papp, K. V., Walsh, S. J. & Snyder, P. J. Immediate and delayed effects of cognitive interventions in healthy elderly: a review of current literature and future directions. *Alzheimers Dement.* **5,** 50–60 (2009).
3. Smith, G. E. *et al.* A cognitive training program designed based on principles of brain plasticity: results from the improvement in memory with plasticity-based adaptive cognitive training study. *J. Am. Geriatr. Soc.* **57,** 594–603 (2009).
4. Thorell, L. B., Lindqvist, S., Nutley, S. B., Bohlin, G. & Klingberg, T. Training and transfer effects of executive functions in preschool children. *Dev. Sci.* **12,** 106–113 (2009).
5. Green, C. S. & Bavelier, D. Action video game modifies visual selective attention. *Nature* **423,** 534–537 (2003).
6. Baddeley, A. D. A three-minute reasoning test based on grammatical transformation. *Psychometric Sci.* **10,** 341–342 (1968).
7. Conklin, H. M., Curtis, C. E., Katsanis, J. & Iacono, W. G. Verbal working memory impairment in schizophrenia patients and their first-degree relatives: evidence from the digit span task. *Am. J. Psychiatry* **157,** 275–277 (2000).
8. Elliott, R. *et al.* Effects of methylphenidate on spatial working memory and planning in healthy young adults. *Psychopharmacology (Berl.)* **131,** 196–206 (1997).
9. Mehta, M. A. *et al.* Methylphenidate enhances working memory by modulating discrete frontal and parietal lobe regions in the human brain. *J. Neurosci.* **20,** RC65 (2000).
10. Owen, A. M., Downes, J. D., Sahakian, B. J., Polkey, C. E. & Robbins, T. W. Planning and spatial working memory following frontal lobe lesions in man. *Neuropsychologia* **28,** 1021–1034 (1990).

11. Sahakian, B. J. *et al.* A comparative study of visuospatial memory and learning in Alzheimer-type dementia and Parkinson's disease. *Brain* **111,** 695–718 (1988).
12. Turner, D. C. *et al.* Cognitive enhancing effects of modafinil in healthy volunteers. *Psychopharmacology (Berl.)* **165,** 260–269 (2003).
13. Owen, A. M. *et al.* Visuospatial memory deficits at different stages of Parkinson's disease. *Neuropsychologia* **31,** 627–644 (1993).
14. Cohen, J. *Statistical Power Analysis for the Behavioral Sciences* 2nd edn (Lawrence Erlbaum Associates, 1988).
15. Gray, J. R., Chabris, C. F. & Braver, T. S. Neural mechanisms of general fluid intelligence. *Nature Neurosci.* **6,** 316–322 (2003).
16. Roca, M. *et al.* Executive function and fluid intelligence after frontal lobe lesions. *Brain* **133,** 234–247 (2010).
17. Wright, S. B., Matlen, B. J., Baym, C. L., Ferrer, E. & Bunge, S. A. Neural correlates of fluid reasoning in children and adults. *Frontiers Human Neurosci.* **1,** 1–8 (2008).
18. Owen, A. M., Doyon, J., Petrides, M. & Evans, A. C. Planning and spatial working memory examined with positron emission tomography (PET). *Eur. J. Neurosci.* **8,** 353–364 (1996).
19. Williams-Gray, C. H., Hampshire, A., Robbins, T. W., Owen, A. M. & Barker, R. A. Catechol *O*-methyltransferase val^{158}met genotype influences frontoparietal activity during planning in patients with Parkinson's disease. *J. Neurosci.* **27,** 4832–4838 (2007).
20. Duncan, J. *et al.* A neural basis for general intelligence. *Science* **289,** 457–460 (2000).
21. Duncan, J., Burgess, P. & Emslie, H. Fluid intelligence after frontal lobe lesions. *Neuropsychologia* **33,** 261–268 (1995).
22. Owen, A. M., Morris, R. G., Sahakian, B. J., Polkey, C. E. & Robbins, T. W. Double dissociations of memory and executive functions in working memory tasks following frontal lobe excisions, temporal lobe excisions or amygdalo-hippocampectomy in man. *Brain* **119,** 1597–1615 (1996).
23. Owen, A. M. *et al.* Frontostriatal cognitive deficits at different stages of Parkinson's disease. *Brain* **115,** 1727–1751 (1992).
24. Wood, S. J. *et al.* Visuospatial memory and learning in first episode schizophreniform psychosis and established schizophrenia: a functional correlate of hippocampal pathology. *Psychol. Med.* **32,** 429–443 (2002).
25. Jäkälä, P. *et al.* Guanfacine and clonidine, alpha 2-agonists, improve paired associates learning, but not delayed matching to sample, in humans. *Neuropsychopharmacology* **20,** 119–130 (1999).
26. Fowler, K. S., Saling, M. M., Conway, E. L., Semple, J. & Louis, W. J. Computerized delayed matching to sample and paired associate performance in the early detection of dementia. *Appl. Neuropsychol.* **2,** 72–78 (1995).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements A.M.O., A.H. and J.A.G. are supported by the Medical Research Council (U.1055.01.002.00001.01 and U.1055.01.003.00001.01). C.G.B. and S.D. are supported by the Alzheimer's Society (UK). We thank the BBC Lab UK team for their contribution, which included the website, task design, data acquisition, recruitment of participants and coordination.

Author Contributions A.M.O. co-designed the study, co-designed the training tasks, designed (with A.H.) the benchmarking tests provided by http://www.cambridgebrainscience.com, co-conducted the statistical analysis, interpreted the data and took overall responsibility for writing each draft of the manuscript. A.H. contributed to the design of the training tasks, designed (with A.M.O) and programmed the benchmarking tests provided by http://cambridgebrainscience.com, co-conducted the statistical analysis and contributed to each draft of the manuscript. J.A.G. co-conducted the statistical analysis, contributed to the interpretation of the data, co-wrote the first draft of the manuscript and contributed to each subsequent version. R.S. designed the data capture, data checking and data cleaning protocols and was responsible for converting data into a format for analysis and for the delivery of the trial database for statistical analysis. He was part of the project management group and contributed to each draft of the manuscript. S.D. contributed to the design of the study, piloted brain training modules, contributed to the design and implementation of the recruitment and retention strategies, was part of the project management group and contributed to each draft of the manuscript. A.S.B. was chair of the independent trial steering committee and advised on key aspects of study design and implementation in this capacity. He also contributed to each draft of the manuscript. R.J.H. advised on key aspects of general study design, contributed to the design of the training tasks and contributed to each draft of the manuscript. C.G.B. jointly conceived of and jointly designed the study and wrote the first draft of the protocol. He was part of the project management group, co-conducted the statistical evaluation, contributed to the interpretation of the data and contributed to each draft of the manuscript.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to A.M.O. (adrian.owen@mrc-cbu.cam.ac.uk).

METHODS

Participants. Of 11,430 participants who met the inclusion criteria, 4,678 were randomly assigned to experimental group 1, 4,014 to experimental group 2 and 2,738 participants to the control group. The relatively reduced number of participants in the control group reflects a greater drop out between the pre-training and post-training benchmarking sessions in this group (equal numbers were assigned to each group at the point of registration), suggesting, perhaps, that the control tasks were less engaging overall than the training tasks. The participants in experimental group 1 completed an average of 28.39 (s.d. = 19.86) training sessions, compared with 23.86 (15.66) in experimental group 2 and 18.66 (12.87) in the control group. The latter result indicates, again, that the control group's task was less engaging than the specific training tasks given to the two experimental groups. In order that 'first' and 'last' scores for performance on the training sessions and the control task could be calculated without error, participants who did not complete at least two training or control sessions between the two benchmarking assessments were excluded from the analysis.

Task design. All three groups were given the same four benchmarking tests twice, once after registering for the trial, but before being shown the training or control tasks, and again six weeks later, irrespective of how many training or control sessions they had chosen to complete in between (subject to the caveat above). The four tests were adapted from a battery of publicly available cognitive assessment tools designed and validated at the Medical Research Council Cognition and Brain Sciences Unit (by A.H. and A.M.O) and made freely available at http://www.cambridgebrainsciences.com. The first test (reasoning) was based on a grammatical reasoning test that has been shown to correlate with measures of general intelligence or *g*[6]. The participants had to determine, as quickly as possible, whether grammatical statements (for example, the circle is not smaller than the square) about a presented picture (a large square and a smaller circle) were correct or incorrect and to complete as many trials as possible within 90 s. The outcome measure was the total number of trials answered correctly in 90 s, minus the number answered incorrectly. The second test (verbal short-term memory (VSTM)) was a computerized version of the 'digit span' task which has been widely used in the neuropsychological literature and in many commercially available brain-training devices to assess how many digits a participant can remember in sequence. The version used here was based on the 'ratchet-style' approach[27] in which each successful trial is followed by a new sequence that is one digit longer than the last and each unsuccessful trial is followed by a new sequence that is one digit shorter than the last. In this way, an accurate estimate of digit span can be made over a relatively short time period. The main outcome measure, average digit span, was the average number of digits in all successfully completed trials. Participants were allowed to make three errors in total before the test was terminated. Versions of the third task (spatial working memory (SWM)) have been widely used in the human and animal working memory literature to assess spatial working memory abilities[8–11,22,28,29]. The version used here[10] required participants to 'search through' a series of boxes presented on the screen to find a hidden 'star'. Once found, the next star was hidden and participants had to begin a new search, remembering that a star would never be hidden in the same box twice. Participants were allowed to make three errors in total before the test was terminated. The main outcome measure was the average number of boxes in the successfully completed trials. The final test (paired-associates learning (PAL)), was based on a task that has been widely used in the assessment of cognitive deterioration in Alzheimer's disease and related neurodegenerative conditions[11,26]. A series of 'window shutters' opened up on the screen to reveal a picture of a different object in each window (for example, a hat or a ball). At the end of each sequence, the participants were shown a series of objects, one at a time, and had to select the correct window for each object. The version used here employed a 'ratchet-style' approach in which each completely successful trial was followed by a new trial involving one more window than the last and each unsuccessful trial was followed by a new trial involving one less window than the last. Participants were allowed to make three errors in total before the test was terminated. The main outcome measure was the average number of correct object-place associations ('paired associates') in the trials that were successfully completed.

During the six-week training period the first experimental group was trained on six reasoning, planning and problem-solving tasks. In the first task (reasoning 1), the participants had to use weight relationships, implied by the position of two see-saws with objects at each end, to select the heaviest object from a choice of three presented below. In the second task (reasoning 2), the objective was to select the 'odd one out' from four shapes that varied in terms of colour, shape and solidity (filled/unfilled). In the third task (reasoning 3), the participants had to move crates from a pile, each move being made with reference to the effect that it would have on the overall pattern of crates and how the result would affect future moves. In the fourth task (planning 1), the objective was to draw a single continuous line around a grid, planning ahead such that current moves did not hinder later moves. In the fifth task (planning 2), the participants had to move objects around between three jars until their positions matched a 'goal' arrangement of objects in three reference jars. In the sixth task (planning 3), the objective was to slide numbered 'tiles' around on a grid to arrange them into the correct numerical order. In all three reasoning tasks and in planning 2, each training session consisted of two 'runs' of 90 s and the main outcome measure was the total number of correct trials across the two runs. For planning 1 and 3, the main outcome measure was the number of problems completed in 3 min.

During the six-week training period the second experimental group was trained on six tests of memory, attention, visuospatial processing and mathematical calculations. In the first task (maths), the participants had to complete simple math sums (for example, 17 − 9) as quickly as possible. In the second task (visuospatial), the objective was to find the missing piece from a jigsaw puzzle by selecting from six alternatives. In the third task (attention 1), symbols (for example, blue stars) would appear rapidly and the participants were required to click on each symbol as quickly as possible, but only if it matched one of the 'target' symbols presented at the top of the screen. In the fourth task (attention 2), the participants were shown a series of slowly moving, rotating, numbers. The objective was to select the numbers in order from the lowest to the highest. In the fifth task (memory 1), the participants were shown a sequence of items of baggage moving down a conveyer belt towards an airport X-ray machine. The number of bags going in did not equal the number of bags coming out. After a short period the conveyor belt stopped and the participant had to respond with how many bags were left in the X-ray machine. In the sixth task (memory 2), the participant was shown a set of cards and asked to remember the picture on each. The cards were then flipped over and the user had to identify pairs of cards with identical objects on them. For all of these tasks, except memory 1, each training session consisted of two 'runs' of 90 s each and the main outcome measure was the total number of correct trials across the two runs. For memory 1, the main outcome measure was the number of problems completed in 3 min.

In each session, the control group were asked five obscure knowledge questions (for example, what year did Henry VIII die?) from one of six general categories (population, history, duration, pop music, miscellaneous numbers and distance) and were asked to place answers in correct chronological order using any available online resource. Each session comprised three sets of five questions and 15 points were awarded for each answer in the correct chronological order.

Data analysis. The main outcome measures were the difference scores (post-training minus pre-training) for the four benchmarking tests in the two experimental groups and the control group in the 'intention to treat' population (that is, those who completed baseline and six-week benchmarking assessments). Comparisons were then made between each of the experimental groups and the control group and between the two experimental groups themselves (Table 2 and Fig. 1). Changes on the training test performance were also calculated by comparing the scores from the first training session with the scores from the final training session.

With such large sample sizes, statistical significance is easily reached, even when actual effect sizes are miniscule, making any numerical differences between two groups very difficult to interpret (as an example, the greater change in VSTM performance in the control group relative to both of the experimental groups is statistically significant, yet is counter to any reasonable hypothesis about brain training and, therefore, has no clear theoretical interpretation). To overcome this problem, the size of any observed differences was quantified by reporting effect sizes together with estimates of the likely margin of error (99% confidence intervals) for all comparisons between groups. Effect sizes provide a measure of the 'meaningfulness' of an effect, with 0.2 being generally taken to represent a 'small' effect, 0.5 a 'medium' effect and 0.8 a 'large' effect[14]. Thus, effect size quantifies the size of the difference between two groups, and may therefore be said to be a true measure of the significance of the difference.

27. Bor, D., Duncan, J., Lee, A. C. H., Parr, A. & Owen, A. M. Frontal lobe involvement in spatial span: Converging studies of normal and impaired function. *Neuropsychologia* **44**, 229–237 (2005).
28. Olton, D. S. in *Spatial Abilities* (ed. Potegal, M.) 325–360 (New York, 1982).
29. Passingham, R. Memory of monkeys (*Macaca mulatta*) with lesions in prefrontal cortex. *Behav. Neurosci.* **99**, 3–21 (1985).